**A blue and white logo

Description automatically generated**

**Bajaj Health Programming Challenge**

**Data Engineering - Qualifier 1**

**Step 1:**

You are given an attendance table that records students’ attendance.

**Table: attendance**

|  |  |  |
| --- | --- | --- |
|  |  |  |

Each student has an attendance\_date entry marked as either 'Present' or 'Absent'.

**Write a python code** to find the students who were absent for **more than three consecutive days**. Your output should include:

* student\_id
* absence\_start\_date (the first day of the absence streak)
* absence\_end\_date (the last day of the absence streak)
* Total absent days in that streak

The above output should **only have the latest absence streak** of the student.

**Expected output from step1:**

A screenshot of a computer

AI-generated content may be incorrect.

**Step 2:**

Using the output from Step 1, write a Python program that:

1. **Join this data** with a students table that contains additional student information.

**Table: students**

A screenshot of a computer

AI-generated content may be incorrect.

1. **Check for valid emails** (A valid email follows the pattern: [something@domain.com](mailto:something@domain.com)).
   1. Email should have @.
   2. The @ character should be followed by domain.com (e.g. @gmail.com is valid, @gmail or @gmailcom is not valid)
   3. Characters before @ **should not have special character (except ‘\_’)** and should not start with number.
2. Add two new columns to the output of step 1:

a. only valid parent emails against the student IDs(Valid email as per bullet point 2) (Column Name in output: **email**)

b. a message in the below format only for students which have a valid parent email (Column name in output: **msg**):

**"Dear Parent, your child [Student Name] was absent from [Absence Start Date] to [Absence End Date] for [Total Absent Days] days. Please ensure their attendance improves."**

1. The code should return the final dataframe as output.

**Final expected output:**

A screenshot of a computer

AI-generated content may be incorrect.

**Expected columns :** ['student\_id', 'absence\_start\_date', 'absence\_end\_date', 'total\_absent\_days', 'email', 'msg']

**Note:**

* **Ensure that the column names are strictly matching with the expected output.**
* **You are requested to edit the run() function in the python file sent to you.**
* **run() function should return pandas dataframe only.**
* **File name should be your ROLLNO.py. For Ex. 12345.py**
* **Submission Link:** [**https://forms.office.com/r/4ctAErYK4B**](https://forms.office.com/r/4ctAErYK4B)
* **Any deviation from the expected output may lead to disqualification.**

**Upload Instructions:**

1. **Instructions to upload a file** 
   1. **Add your file in github repository.**
   2. **Go to the uploaded file.**
   3. **Click on raw button.**
   4. **Copy the url and submit it in the form.** 